A NEW APPROACH TO PILOT RATING SCALES

Coral Vance Schufeldt

United States
Naval Postgraduate School

THESIS

A NEW APPROACH TO PILOT RATING SCALES

by

Coral Vance Schufeldt

Thesis Advisor: D. M. Layton

September 1971

A New Approach to Pilot Rating Scales

by

Coral Vance Schufeldt
Lieutenant Commander, United States Navy
B.S., United States Naval Academy, 1963

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN AERONAUTICAL ENGINEERING

from the

NAVAL POSTGRADUATE SCHOOL
September 1971

ABSTRACT

An examination and comparison of pilot rating scales presently in use and an investigation into the possibilities of a linear rating scale were conducted. The hypothesis was advanced that a rater may transpose his impression of performance directly to a non-adjectival, non-ordinal rating scale and thereby relate his psychological continuum to a numerical index. Experimental data, though limited, tended to support this hypothesis.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

With the advent of flight vehicles with operating envelopes ranging from terra firma to the threshold of space and beyond, the environmental and dynamic spectrums encountered on a single flight are all-encompassing. Man is the low frequency response component [Ref. 1] in the overall closed-loop man-machine system, therefore, control systems must be designed within manageable limits. In short, the effort expended in vehicle control must be minimized so that the pilot may be free to complete other duties in the cockpit.

Consequently, the suitability of a machine system to serve its intended mission is ultimately determined by a series of evaluations. The most difficult of these assessments occurs at the man-machine system interface.

Pilot evaluation of handling qualities determines the suitability of the machine system, yet there remains to be found a set of universally acceptable parameters for this evaluation. The complete nature of a pilot's task, work load, mental stress and acuity have not been described in any form of analytically determined transfer function or performance index [Ref. 2]. It is assumed, however, that there exists a relationship between pilot comment and performance and/or vehicle handling qualities.

Efforts to standardize the qualitative aspects of language into a quantitative handling quality rating have been made. It is the purpose of this study to examine and compare the rating scales

presently in use and to investigate the possibilities of a linear rating scale with its inherent advantages.

II. HISTORY

A. EARLY DEVELOPMENTS

During the early 1930's when aviation was maturing, the need to delineate acceptable aircraft parameters was recognized. Consequently, a "check list" for this purpose was proposed by Edward P. Warner [Ref. 3]. Subsequent work by Soule' and by R. R. Gilruth at the Langley Laboratory of NACA condensed these requirements and a set of specifications for military aircraft acceptance eventually resulted [Ref. 4].

After this initial break-through in establishing aircraft specifications, emphasis was placed on devising pilot opinion ratings aimed at specific problem areas. The concept of a general pilot rating received little attention.

B. COOPER SCALE

In 1957 at the annual meeting of the Flight Testing Session, Institute of Aeronautical Sciences, Ames' Chief Research Pilot George E. Cooper introduced a generalized pilot rating scale which enjoyed immediate and almost total acceptance [Ref. 5]. This epoch scale (Fig. 1) synthesized the previous work of NACA Langley and thereby provided an authenticated scale which could be applied to any aircraft handling qualities evaluation. It was the first rating scale to associate the qualitative nature of pilot opinion with a quantitative index.

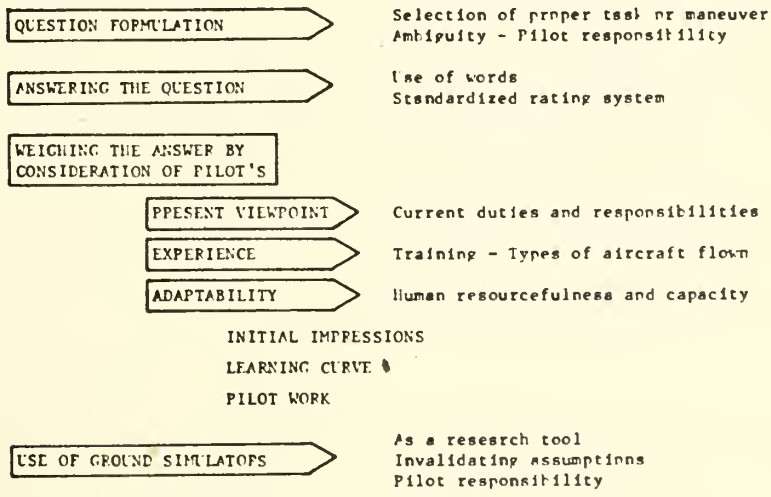
In applying this scale, it was recommended that the evaluator pay particular attention to question formulation (Fig. 2). The

	ADJECTIVE RATING	NUMERICAL RATING	DESCRIPTION	PRIMARY MISSION	CAN BE LANDED
NORMAL OPERATION	Satisfactory	1	Excellent, includes optimum	Yes	Yes
		2	Good, pleasant to fly	Yes	Yes
		3	Satisfactory, but with some mildly unpleasant characteristics	Yes	Yes
EMERGENCY OPERATION	Unsatisfactory	4	Acceptable, but with unpleasant characteristics	Yes	Yes
		5	Unacceptable for normal operation	Doubtful	Yes
		6	Acceptable for emergency condition only*	Doubtful	Yes
NO OPERATION	Unacceptable	7	Unacceptable even for emergency condition*	No	Doubtful
		8	Unacceptable - Dangerous	No	No
		9	Unacceptable - Uncontrollable	No	No
	Unprintable	10	"Motions possibly violent enough to prevent pilot escape"		

*Failure of stability augmenter

COOPER SCALE

FIGURE 1



FACTORS AFFECTING PILOT OPINION

FIGURE 2

question had to be sufficiently specific so as to minimize interpretation and ambiguity.

The pilot, in answering the question, was required to channel his exposure, sensations and reactions into the scale vocabulary by first considering four handling qualities categories: Satisfactory, Unsatisfactory, Unacceptable and Unprintable. As may be noted from Figure 1, these categories were separated, for description purposes, at the approximate values 3.5, 6.5 and 9.5 respectively. Within each category, the pilot was required to further define his opinion in terms of the scale vocabulary and a secondary mission (landing).

Once the pilot had formulated his opinion with respect to the scale, his evaluation had to be weighted in consideration of his viewpoint, experience and adaptability. For example, a patrol pilot might evaluate the stall-associated buffet and departure in a fighter as "Unacceptable-Dangerous" (numerical rating 8); whereas, a fighter pilot might evaluate the same characteristics as "Satisfactory, but with some unpleasant characteristics" (numerical rating 3). Then, with some exposure, the same two pilots might reevaluate the characteristics at 4 and 2 respectively. The rating scale was, therefore, very subject to experience and adaptability. To eliminate this deficiency and to provide some measure of consistency, it was suggested that the scale be used only by test pilots.

Though the Cooper Scale had claim to primacy, it was ambiguous in its definitions and complicated in that it placed stipulations on pilot opinion. It would appear that the scale was designed to evaluate aggregate handling qualities.

C. HARPER SCALE

Robert P. Harper, Jr. used a pilot opinion scale (Fig. 3) for evaluating the handling qualities of a variable stability aircraft in 1959 [Ref. 6]. The Harper Scale was developed honoring the stipulations of question formulation but with a concept quite different from the Cooper Scale. Harper was interested in evaluating pilot-vehicle performance, but found this extremely difficult because of pilot adaptability. Instead, a scale was devised to evaluate pilot opinion with respect to alterations in the stability derivatives and thereby arrive at a pilot preference: a most suitable aircraft stability.

To ensure reliability and compensate for scale vocabulary deficiencies, test pilots wire-recorded their subjective comments during the evaluation and recorded their scale rating following each evaluation. This was, perhaps, the best aspect of the testing procedure. The pilot rating was kept simple and subordinate to the subjective evaluation. Because of this reliance on subjective comments made during the tests, the pilot rating was utilized as a cursory index to the evaluation and not as an end in itself.

In evaluating the handling qualities with respect to the rating scale, the pilot considered four handling qualities categories: Acceptable and Satisfactory, Acceptable but Unsatisfactory, Unacceptable, and Unflyable. The separation between these categories occurred at 3.5, 6.5 and 9.5 respectively. Within each category, the pilot further defined his opinion in terms of a single, though sometimes ambiguous, adjective (Fig. 3).

CATEGORY	ADJECTIVE DESCRIPTION WITHIN CATEGORY	NUMERICAL RATING
Acceptable and Satisfactory	Excellent Good Fair	1 2 3
Acceptable but Unsatisfactory	Fair Poor Bad	4 5 6
Unacceptable	Bad* Very bad** Dangerous***	7 8 9
Unflyable	Unflyable	10

*Requires major portion of pilot's attention
**Controllable only with a minimum of cockpit duties
***Aircraft just controllable with complete attention

HARPER SCALE

FIGURE 3

D. HARPER SCALE ADAPTATIONS

In contrast to the Cooper Scale, the Harper Scale (often cited as the Cornell or CAL Scale because of its extensive use by Cornell Aeronautical Laboratory, Inc.) was designed as an index for evaluating particular and highly restricted handling qualities. Efforts to adapt the CAL Scale to the evaluation of aggregrate handling qualities met with varied success.

One such example was the application made by Michael L. Parrag in 1967 [Ref. 7] in studying the effects on handling qualities of higher-order response characteristics against a background of varying conditions and associated mission tasks.

To facilitate more reliable and consistent pilot comments, the test pilots were provided with a comment check list for the two flight conditions (Fig. 4), and instructed to make subjective comments following each test run. After all tasks were completed, a comprehensive subjective report was required incorporating all the salient features of each configuration. Finally, an objective report using the comment check list was made.

Here, as in Ref. 4, emphasis was placed on subjective comments. Task-oriented objective comments were used to provide consistency and point out features of each task which might otherwise have been overlooked. Although the CAL Scale was used as an index to pilot opinion, it was, for all practical purposes, insignificant in evaluating the handling qualities investigated.

E. COOPER-HARPER SCALE

With wide and independent usage of the Cooper and Harper Scales

APPPOACH COMMENT CARD

1. IS THE AIRPLANE DIFFICULT TO TRIM?
2. IS ATTITUDE CONTPOL SATISFACTOPY?
   TENDENCY TO OVERCONTROL?
3. IS MAINTAINING ALTITUDE A PPOBLEM?
   a) STRAIGHT AND LEVEL
   b) TURNS
4. IS MAINTAINING AIRSPEED A PROBLEM?
5. WERE GLIDE SLOPE ERRORS EASILY CORRECTED?
   WAS IT DIFFICULT TO MAINTAIN GOOD GLIDE SLOPE CONTPOL?
6. WHAT INSTRUMENTS ARE YOU USING MOST?
7. COULD YOU MAKE AN INSTRUMENT LANDING APPPOACH WITH THIS CONFIGURATION AT THE SPEED OF 125 KNOTS?
8. PILOT RATING - ADJECTIVES - NUMBER - WHY?

HIGH ALTITUDE COMMENT CARD

1. IS THE AIRPLANE DIFFICULT TO TRIM?
2. IS ATTITUDE CONTROL SATISFACTORY? TENDENCY TO OVER-CONTROL?
3. IS NORMAL ACCELERATION CONTROL A PROBLEM?
4. IS HOLDING ALTITUDE A PROBLEM?
   a) STRAIGHT AND LEVEL
   b) TURNS
5. ARE THERE ANY DIFFICULTIES IN FLIGHT PATH CONTROL DURING THE CLIMBING AND DESCENDING TURNS?
6. ARE THERE ANY PPOBLEMS ASSOCIATED WITH THE TRACKING TASK?
7. PILOT RATING - ADJECTIVES - NUMBER - WHY?

PILOT COMMENT CARDS

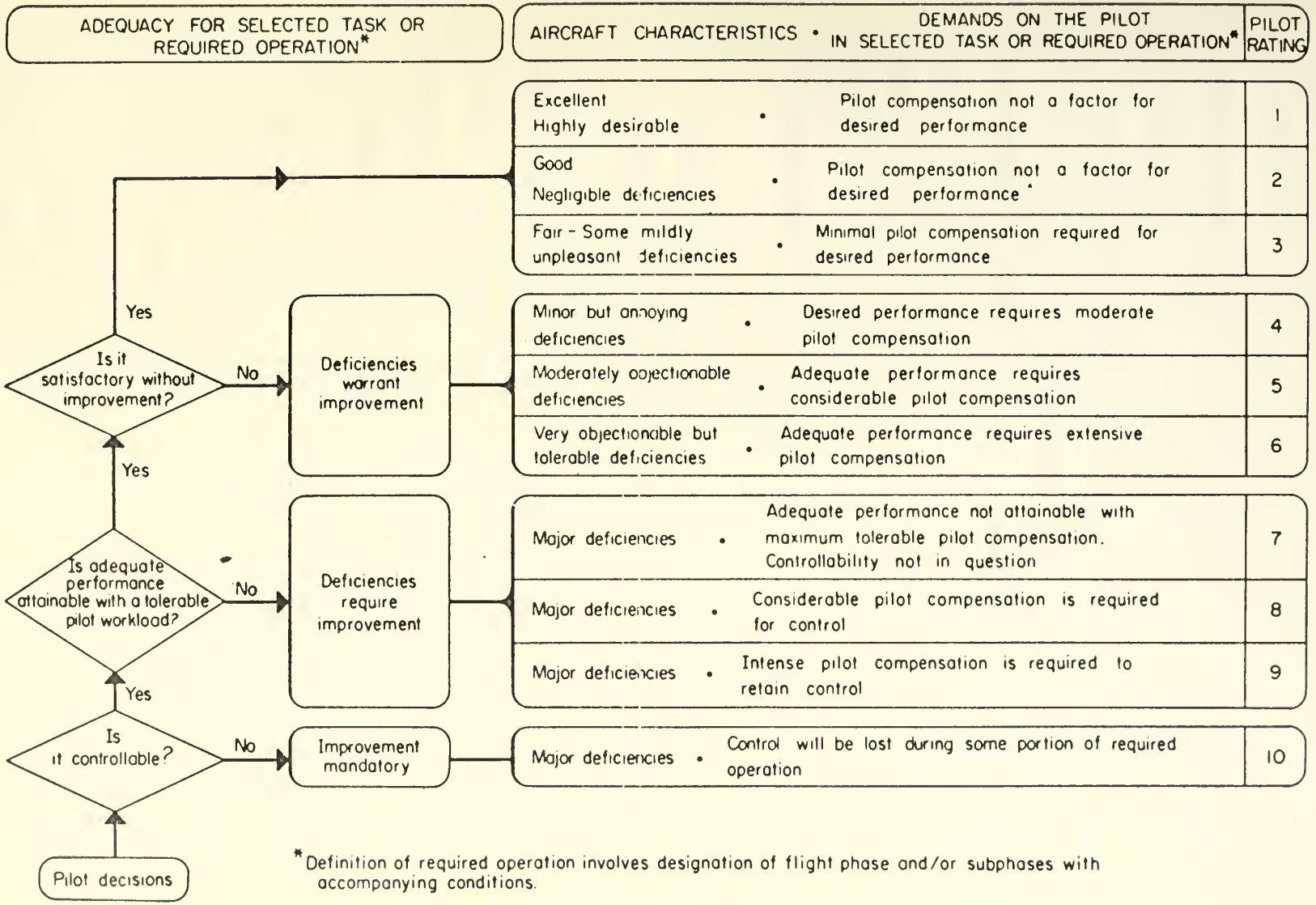
FIGURE 4

the problems previously cited for each were sources of confusion in application. It became increasingly apparent that an acceptable composite rating system incorporating the best features of each scale would be advantageous.

To this end Cooper and Harper jointly advanced a revised rating scale in 1966 [Ref. 8]. This scale (Fig. 5), hereafter referred to as the Cooper-Harper Scale, enjoyed general acceptance and preference over the previous scales; however, the various implementing institutions voiced a need for clarification in semantics and in application. In 1969 an explicitly comprehensive joint report was published to modify and clarify the Cooper-Harper Scale [Ref. 9]. The report precisely defined flight evaluation terminology and discussed the aspects of question formulation and scale data application.

Based on the voluminous data and comments available from international audiences of the Cooper and Harper Scales, the Cooper-Harper Scale was excellently designed as a dichotomous procedure of evaluation. A pilot, in evaluating a handling quality, systematically chose between two alternatives which channeled his consideration into a rating category or into another dichotomous decision with the same channeling result. Through this simplified procedure (compare with the relative complexity of previously discussed procedures) three of four existing categories were eliminated without ever considering the applicable descriptive adjectives.

The inverted ten-point scale was retained in the interests of consistency. An ordinal sequence varying in magnitude with the degree of "goodness" would seem more appropriate; however, audiences of the previous scales had become accustomed to the



*Definition of required operation involves designation of flight phase and/or subphases with accompanying conditions.

COOPER-HARPER SCALE

FIGURE 5

inverted scale and a reordering of the numerical indices would have resulted in unnecessary confusion. To further ease the transition from previous scales, the boundaries of 3.5, 6.5 and 9.5 were retained.

It would appear that a satisfactory method for assessing the man-machine interface had been achieved; but not quite. Although the Cooper-Harper Scale continues to be the most widely used evaluation system to date, it remains insensitive at the bad end and does not exhibit the desirable feature of linearity. Linearity is that feature of a rating scale which will allow the averaging of data ensembles without distorting the data sample interpretation.

F. McDONNELL SCALE

In 1968, J. D. McDonnell published his study of rating techniques [Ref. 10]. His objective was to evolve a rating scale which had an underlying linear structure to facilitate mathematical operations on pilot data. This underlying structure required the discipline of psychophysics for determination.

Although a detailed examination of psychophysics is beyond the scope of this study, the basic theory is presented for clarification. If an objective measure is made upon some object, the resulting data must lie along some physical continuum. If an evaluator estimates a measure, the measure is subjective and must lie along some psychological continuum. The relationship between these two continua, if it could be determined, would provide a means of linearizing the subjective scale.

To establish an intervaled psychological continuum, a list of sixty-four appropriately descriptive phrases were randomly submitted to sixty-three raters. For each phrase, the raters were instructed to indicate their impression of a hypothetical vehicle so described on a plot with the end points of "most favorable" and "least favorable". The data were then processed by the methods of psychophysics and successive intervals and assigned a relative standing on a scale of nine. The data were further reduced to the arbitrary seven-point McDonnell Scale depicted in Figure 6.

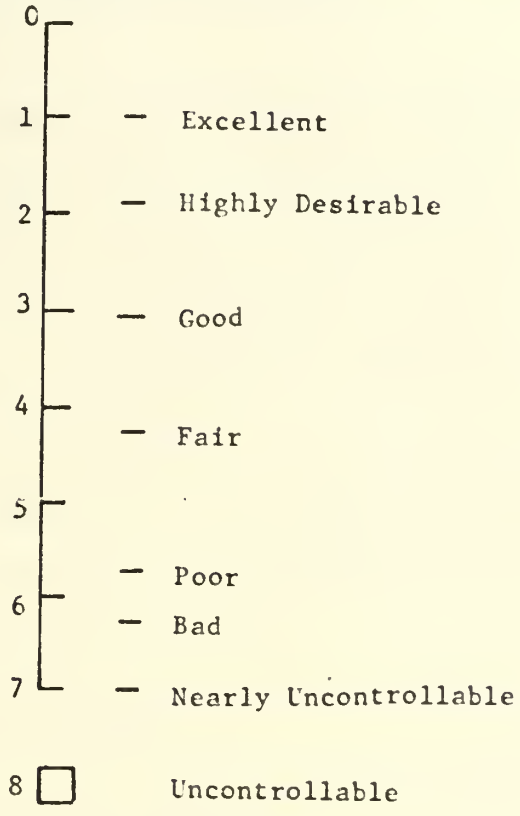
The McDonnell Scale (often called the Global Rating Scale because it related aggregate handling qualities) was, therefore, presumed to be a linear scale of constant subjective sensitivity reflecting the resolving power of raters to distinguish semantic differences. Because it was related to a seven-point scale in contrast to the ten point scales with which users were familiar, it was not accepted with any noticeable exuberance.

The truely important contribution made by McDonnell was the list of evaluation phrases related to an index of nine and reflecting psychological sensitivity. The phrases were divided into six categories: Handling Qualities, Control, Precision, Response Characteristics, Effects of Deficiencies, and Demands on Pilot. Through the use of this listing, specialized linear scales may be constructed to satisfy particular rating requirements.[1]

---

[1] See APPENDIX B.

FAVORABILITY OF HANDLING QUALITIES



MCDONNELL SCALE

FIGURE 6

G. CONTEMPORARY RESEARCH

In designing the washout circuitry[2] for the Ames All-Axis Motion Generator, it became necessary and expedient to solicit pilot opinion in determining the "best" set of parameters to use in a given configuration. To this end, S. F. Schmidt and Bjorn Conrad [Ref. 11] used three non-ordinal, relative rating scales in their evaluations (Fig. 7-a, b, c).

The questions related to each scale were particularly tailored to the descriptive adjectives shown and they were simple in nature. By using pilot comments as an index, the design providing the best overall simulator characteristics was obtained. However, moderate changes in the washout circuitry initially selected did not alter pilot opinion during subsequent testing.

It would appear that one or both of the following factors were responsible for the inability of rating pilots to distinguish minor changes in simulator characteristics:

1. The evaluation task was insensitive to minor changes in system response
2. The rating scale adjectives were too widely separated on psychological continuum.

During a personal interview[3] Conrad discussed the work on which he had reported in Ref. 11. In determining the best washout circuitry the pilot ratings extracted from his scales were heavily supplemented

---

[2] That servo circuitry of an all-axis motion simulator which provides for returning the simulator to its initial position after being disturbed. It is important that this function be executed at a rate below a pilot's sensing threshold.

[3] Interviewed on 10 May 1971 at Analytical Mechanics Associates, Inc. of Palo Alto, California.

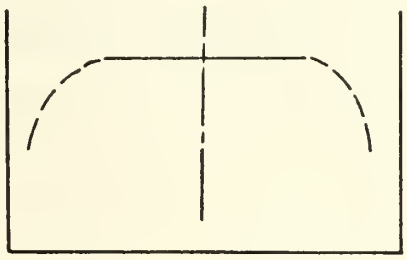
☐ EXCELLENT	☐ MORE DIFFICULT
☐ GOOD	☐ SLIGHTLY MORE DIFF.
☐ FAIR	☐ ABOUT THE SAME
☐ POOR	☐ LESS DIFFICULT
☐ UNACCEPTABLE	☐ SUBSTANTIALLY EASIER
7-a	7-b
☐ ALWAYS	☐ MUCH HARDER
☐ OFTEN	☐ HARDER
☐ OCCASIONALLY	☐ SAME AS
☐ RARELY	☐ EASIER
☐ NEVER	☐ MUCH EASIER
7-c	7-d

CONRAD SCALES

FIGURE 7

with debriefs. It was primarily through this method of pilot interview that the best washout circuitry was obtained.

He observed that pilots rapidly adapted to minor configuration changes without altering their rating, and he described this lack of sensitivity as a rating plateau (Fig. 8).



Performance
PILOT RESPONSE

Figure 8

He additionally noticed that a pilot's impression of his mean performance changed from day to day. This, therefore, required that at least one test run utilizing the "standard" washout circuitry be conducted to reestablish the pilot's mean performance, a time-consuming and costly procedure.

Conrad's present work, an extension of that above, tasks pilots with flying formation on the television display of a six-degree of freedom simulated tanker aircraft. It is hishope that this relative position task will prove to be sufficiently sensitive and thereby provide reliable pilot ratings on the scale depicted in Figure 7-d.

H. SUMMARY

The rating scales which have been reviewed fall into the two categories, as distinguished according to purpose, of aggregate and

relative handling qualities evaluations. The first category consists of the Cooper and Cooper-Harper Scales; whereas, the latter consists of the Harper, McDonnell and Conrad Scales.

During a personal interview[4] Cooper related the circumstances stimulating the evolution of his Scale. While evaluating a variable stability F6F Wildcat the project engineers had an understandable tendency to mathematically manipulate the flight data in the course of its reduction; however, the conclusions derived therefrom did not necessarily reflect the pilot's interpretation of the actual handling qualities encountered. To eliminate this inadvertent misinterpretation of flight data, the Cooper Scale was designed.

When Cooper presented his Scale at the annual meeting of the Institute of Aeronautical Sciences it was immediately accepted and internationally implemented as an aggregate evaluation scale. Though the Cooper Scale was not designed for this purpose, international usage determined its application.

In the collaborative effort to develop the Cooper-Harper Scale, Harper advocated a relative evaluation scale; however, the various implementing institutions preferred a scale applicable to aggregate evaluations and the dichotomous scale resulted.

The Harper and Conrad Scales were obviously designed to evaluate relative handling qualities and no further discussion is necessary.

The McDonnell (or Global) Scale was designed as an aggregate rating scale; however, because of its syntactical simplicity it could

---

[4]Interviewed on 10 May 1971 at the Ames Research Laboratory, NASA, NAS Moffett Field, California.

be applied only to relative evaluations (see Fig. 6). The sixty-three psychologically intervaled phrases resulting from McDonnell's research, however, were applicable to both aggregate and relative handling qualities evaluations.

In evaluations utilizing any of the rating scales except the Cooper-Harper Scale subjective pilot comment was required to provide meaningful evaluation data.

III. HUMAN RESPONSE

A. INTRODUCTION

The Cooper-Harper Scale was excellently designed and remains the best aggregate rating scale in existence because of its dichotomous nature and its acceptance as the international standard. However, it was specifically designed so as not to facilitate the averaging of ratings [9].

With the advent of greater sophistication in aircraft research and development, it has become increasingly important to evaluate the relative "goodness" of aircraft components and subsystems. It is assumed that a highly desirable aerospace vehicle may be designed and built; however, a rating scale capable of reliably determining the acceptance or rejection of one highly desirable system over another is yet to be evolved. It is the purpose of this section to investigate the possibility of such a rating scale.

For a scale to effectively reflect minor differences in performance, extreme sensitivity is desired. The inherent advantages of linearity are also desired to facilitate mathematical operations on a limited ensemble and thereby suppress research and procurement costs.

The hypothesis of this investigation is that a linear rating scale coincident with the psychological continuum begets sensitivity. The psychological continuum was investigated [10] and resulted in the McDonnell Scale, but, as may be noted from Figure 6, descriptive adjectives and/or phrases did not align cardinally. This, then, provided a source of confusion because the numerical value associated

with the adjective might not coincide with the rater's psychological continuum. Were this source of syntactic confusion eliminated, the rater could transpose his impression of performance directly to a rating scale and thereby relate his psychological continuum to a linear numerical index. Additionally, if allowed to fractionalize his rating, sensitivity would be limited only by the rater's discriminate dispersion[5] and frustrations.

To investigate this hypothesis, a simple puzzle was selected and submitted to the analytically inclined students in the Department of Aeronautics of the Naval Postgraduate School. Upon completion of the test, or at the expiration of an allotted time, the subjects were asked to rate their impression of the difficulty they encountered in working the puzzle on three numerical scales.

B. TEST EQUIPMENT

The plastic Kohner EVEN-STEVEN solitaire puzzle (Fig. 9) was used as the testing device. It consisted of a base with eight equal depth holes, eight equal length sleeves with variable interior depths, and eight variable length pegs. The puzzle had 40, 320 (eight factorial) different solutions, one of which resulted in all pegs being even.

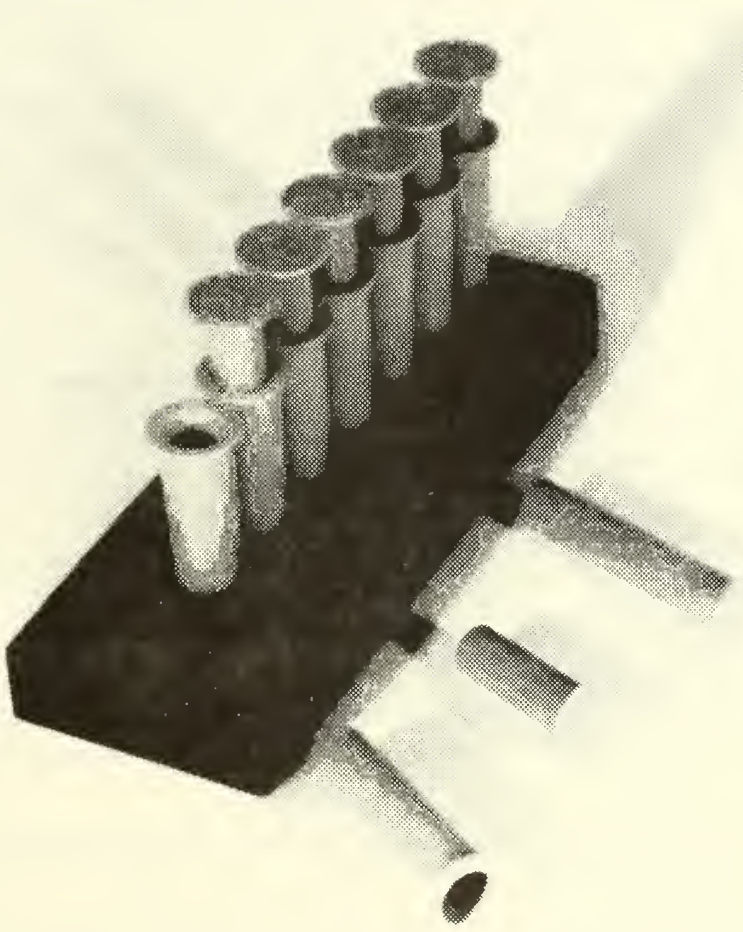
A standard stop-watch was used for timing, and the scales depicted in TABLE 1 were used for rating purpose.

C. TESTING PROCEDURE

Before starting the exercise, the subjects were briefed in detail regarding the physical characteristics of the puzzle. Prior to each

---

[5]The deviation of the resolving power of raters to distinguish minor differences in performance.



EVEN-STEVEN

FIGURE 9

TABLE I

RATER QUESTIONAIPRE

NAME _____ DATE _____

AGE _____

You are requested to solve the EVEN-STEVEN puzzle as a Human Response Section of a Thesis. You will have 60 seconds in which to complete the exercise. After completing, please indicate the degree of Difficulty you encountered while performing the exercise.

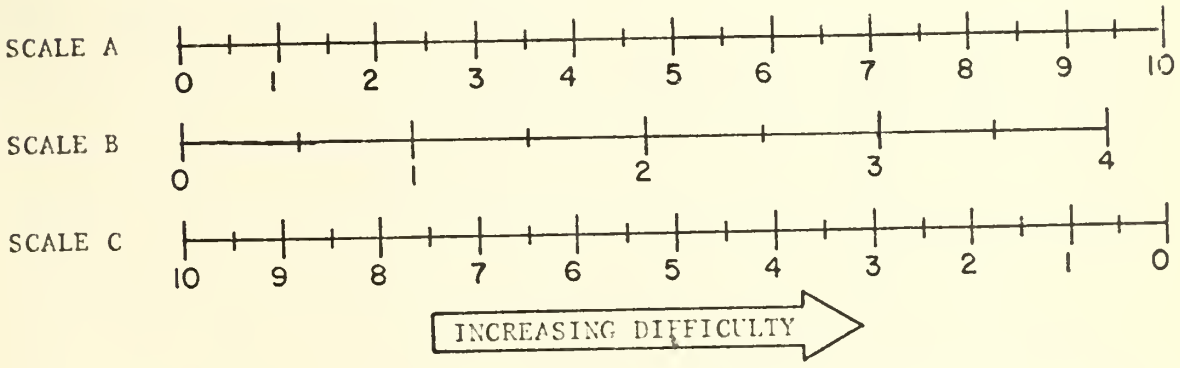
DIRECTIONS:

1. Set the purple sleeves in the base.
2. Insert the orange pegs in the sleeves so that the tops are even (make STEVEN-EVEN).
3. Do not look into the sleeves for distance estimations.

TEST: 60 seconds

RATING: Below are three rating scales with the direction of Increasing Difficulty as indicated. Use all three scales.

1. Indicate your impression of Difficulty in the box provided.
2. Indicate the scale you prefer; A, B or C.



	TEST 1			TEST 2			TEST 3		
	A	B	C	A	B	C	A	B	C
RATING									
SCALE PREFERENCE									
SCORE									
TIME									

test the pegs and sleeves were removed from the base and mixed randomly within a box before the subject. The exercise was started on the proctor's "mark" with the subject's hands poised over the box. At test completion the time was recorded or, if the subject did not complete the test in 60 seconds, the number of even pegs, regardless of height, was recorded. The elapsed time or number of even pegs was the basis for determining performance.

The subject was then asked to rate his impression of the difficulty he encountered in working the puzzle with respect to all three scales on the RATER QUESTIONNAIRE (TABLE 1), and to indicate his rating in the box provided. This procedure was repeated twice to provide for three tests. When subjects inquired as to the degree of difficulty associated with scale end points, they were briefed that this determination was the rater's responsibility. By so doing, the rater's personal psychological continuum was enjoined.

D. RESULTS AND DISCUSSION

The exercise was administered to thirty-one subjects as outlined above, and the raw data were recorded in Appendix A. Of the subjects tested, 25 or 80.8% understood the rating procedure. The remaining six failed to rate their impression of the difficulty they encountered as evidenced by their constant ratings on each scale, regardless of their performance, throughout the testing sequence. Consequently these data were discarded because it was impossible to determine the linear correlation of a point.

1. Question Formulation and Interpretation

The failure of 19.2% of the subjects to comprehend the rating procedure may be the result of incorrectly written rating statements

(i.e., "...please indicate the degree of Difficulty you encountered while performing the exercise." and "Indicate your impression of Difficulty in the box provided.") However, these statements were combined during the pretesting brief (i.e., "Indicate your impression of the Difficulty you encountered in working the puzzle.").

When this 19.2% was queried regarding their constant rating, all replied that the difficulty of the test was a constant regardless of their performance.

Subjects 7, 12 and 14 (TABLE II) all had inappropriately low correlation factors because their ratings indicated increased difficulty for increased performance. When each was queried, he related that more incorrect puzzle combinations were discovered in subsequent testing; consequently, his impression of puzzle difficulty increased. Although these ratings did not properly reflect the rating statements, they were used in the Linearity section because such deviations may be expected in any testing procedure.

2. Linearity

Linear correlation assumes a linear relationship between variables. If a series of variables are linearly related, the correlation factor will be 1.00. Deviations from linearity will yield factors less than 1.00.

To facilitate detailed analysis and to justify raw data averaging, an individual correlation factor (r) was calculated for each exercise subject listed in TABLE II. In correlation factor calculations the time to exercise completion or the number of even pegs was used as the independent variable, and the subject's rating was used as the dependent variable.

TABLE II

RATER PERFORMANCE - RATING
CORRELATION FACTOR

CORRELATION FACTORS (r)

SUBJECT	SCALE A	SCALE B	SCALE C	SUBJECT	SCALE A	SCALE B	SCALE C
1	.992	.993	.993	14	.143	.142	.142
2	.995	.983	.956	15	.986	.999	.986
3	.992	.997	.992	16	.901	.998	.998
4	.905	.905	.905	17	.999	.971	.999
5	.993	.993	.993	18	.999	.982	.929
6	.899	.739	.897	19	.999	.499	.999
7	.181	.181	.181	20	.866	.866	.866
8	.938	.939	.939	21	.960	.961	.961
9	.596	.659	.596	22	.997	.998	.998
10	.866	.831	.866	23	.997	.953	.976
11	.545	.645	.600	24	.999	.999	.999
12	.189	.346	.453	25	.997	.993	.968
13	.997	.999	.999				

Scales A and B yielded correlation factors of which 90.9% were greater than 0.8 and 81.8% were greater than 0.9. Scale C yielded 77% and 72% respectively. The overall correlation factors for Scales A, B and C were 0.928, 0.905 and 0.927 respectively. This high degree of performance-rating correlation confirmed linearity and sensitivity, and was an extremely strong indication that raters were able to relate their personal psychological continuum to a linear, non-adjectival, non-ordinal rating scale. It additionally provided justification for the averaging of ratings.

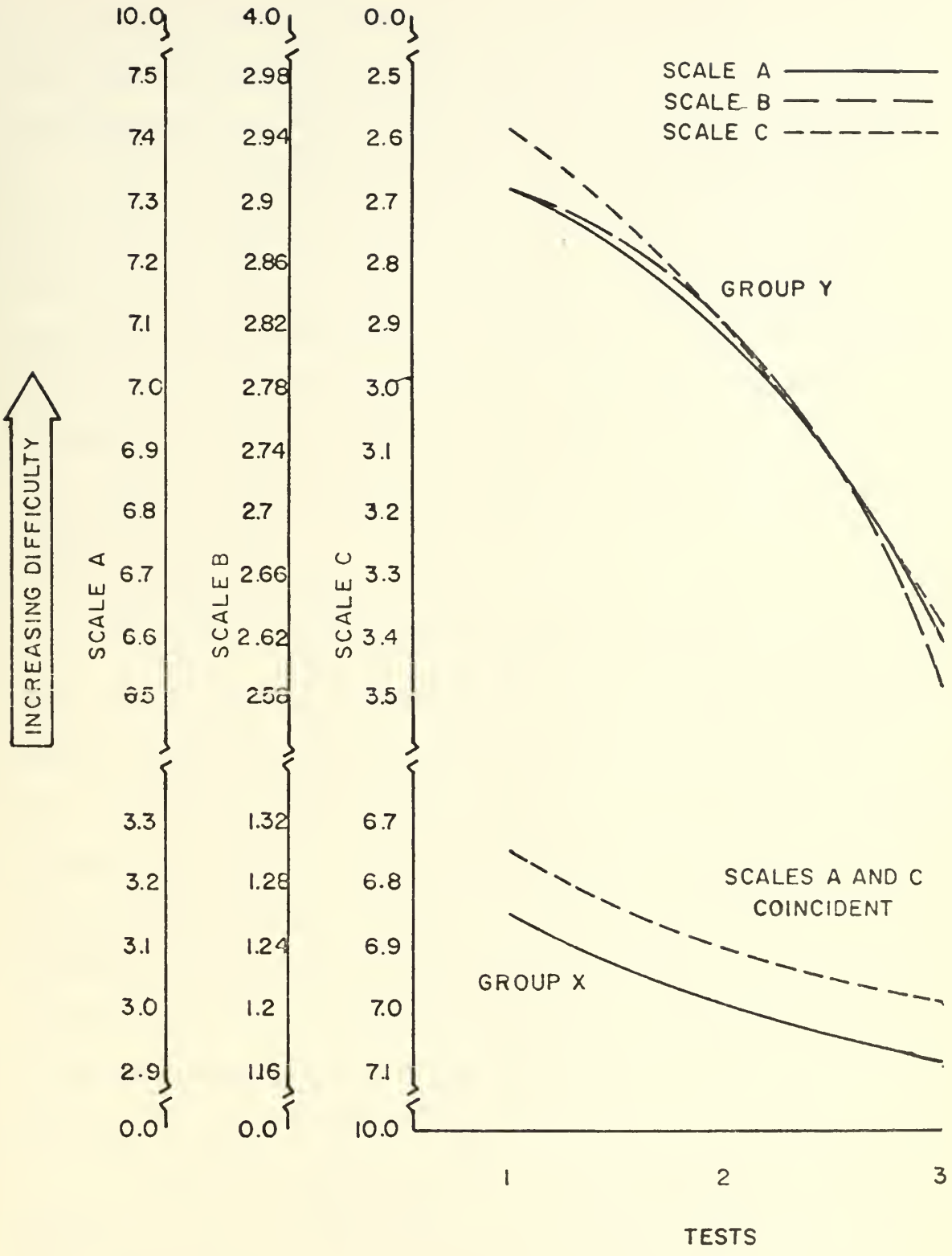
Another feature of high correlation is that relatively few trials may be conducted with a high degree of confidence in the resulting data. This thereby reduces the time and cost expenditures associated with testing.

3. Rating Analysis

The test subjects' ratings fell into two groups as characterized by those who completed all tests during the allotted time (Group X) and those who completed two or less tests (Group Y). As indicated in Figure 10, Group X experienced less difficulty than Y throughout the testing sequence; however, the rating curves of Group X reflected decreased learning in contrast to the curves of Group Y.

It should be noted that the rating curves of Group Y did not remain parallel as did those of Group X. This was, perhaps, an indication of the frustration experienced in not being able to complete each test. Such a factor would influence rating accuracy and, consequently, rating sensitivity.

By averaging the unweighted corresponding test ratings of both Groups (there were more subjects in Group Y), Figure 11 was



GROUP RATING CURVES

FIGURE 10

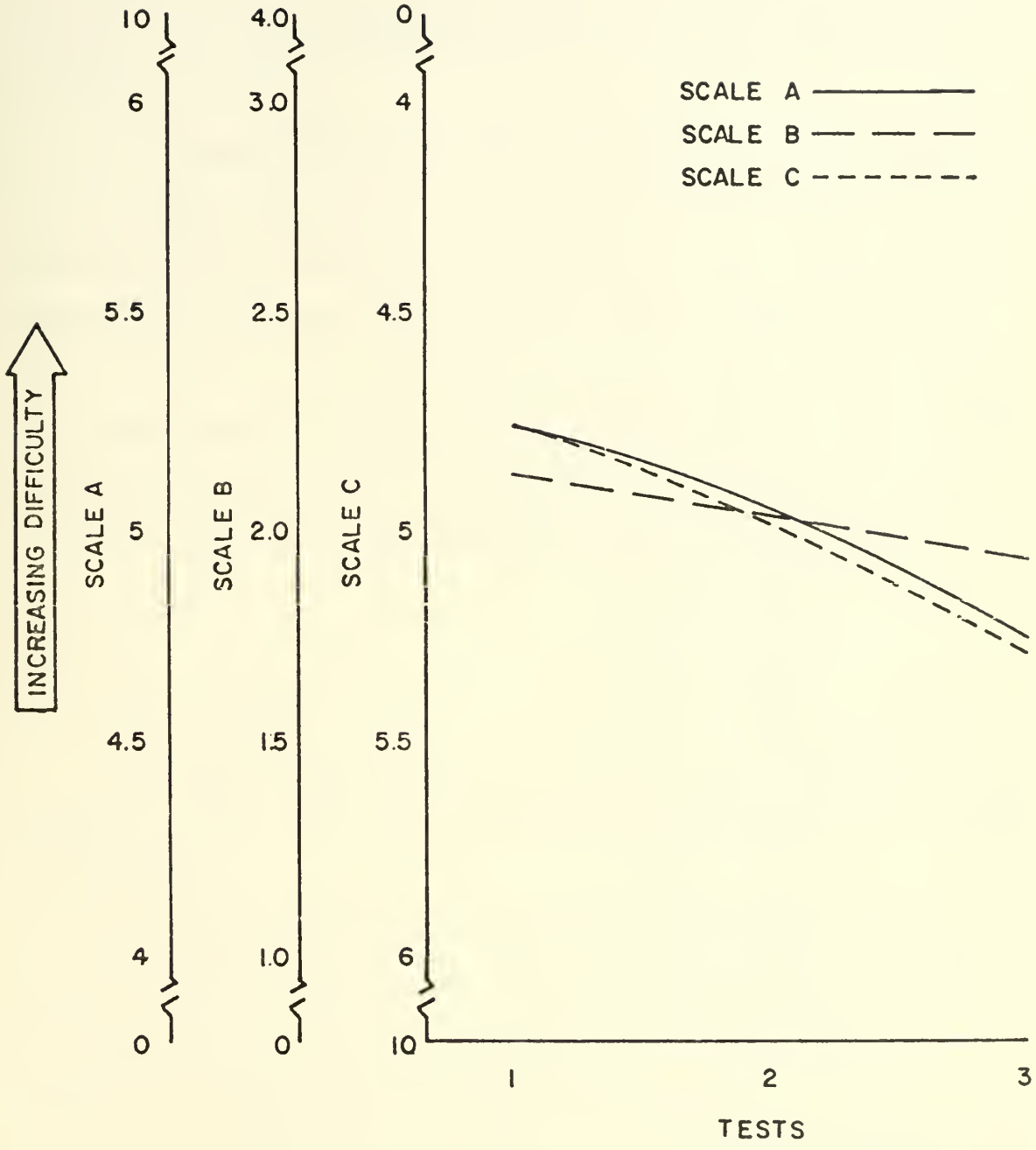
constructed. As may be observed, the average rating curves ranged about the numerical mean of each scale, and, in fact, the average ratings of Scales A, B and C were 5.00, 2.02 and 5.02 respectively.

Considering these two facts, it must be assumed that the test subjects discarded any "degree of difficulty" associated with the scale end points and related all of their ratings to the scale numerical mean. Consequently, all rating was a matter of judgment; a matter of relating their psychological continuum to the scales presented in TABLE I. Whether test subjects consciously or subconsciously related to the scales' numerical means was beyond the scope of this study.

4. Scale Preference

Of the 31 test subjects, 28 preferred Scale A, two preferred Scale B, and one preferred Scale C. It was interesting to note that Scale A construction paralleled that of the Cooper-Harper Scale (i.e., increasing numerical index with increasing degree of "badness"); however, only 35% of the test subjects had ever been exposed to the Cooper-Harper Scale. Because the subjects were enrolled in a mathematically oriented curriculum, the preference for a decimal system based on ten seemed appropriate. As evidenced from the overall correlation factors and the Scale average ratings, the preference for Scale A appeared valid.

The limited preference for Scale B was believed to reflect exposure to the 4.0 Navy system. The preference for Scale C was believed to have been made in the interests of inconsistency and levity.



AVERAGE RATING CURVES

FIGURE II

5. Recommendations

Because of the high correlation experienced during this investigation and the preference of raters for a decimal scale based on ten, it is recommended that such a scale be used in all relative rating evaluations.

Although the Cooper-Harper Scale is unequivocally accepted for its designed purpose, it could be improved if the scale advanced herein were used to evaluate the relative "goodness" within a Cooper-Harper ordinal category. For example, once an ordinal category were determined via the dichotomus procedure, the category could be further defined by Scale A utilization. To designate such a refinement, the first number of a series could reflect the non-averagable Cooper-Harper rating and subsequent numbers reflect Scale A (i.e., 1.2.25).

E. CONCLUSIONS

The purpose of this study was to examine and compare the rating scales presently in use and to investigate the possibilities of a linear rating scale.

A review of rating scale development and a study of the current rating scales were presented in Section II. Section II also provides an organized source of information for the rating-scale novice that may be used to develop specialized rating scales.

Section III advanced with some substantiation the hypothesis that a rater may transpose his impression of performance directly to a non-adjectival, non-ordinal rating scale and thereby relate his psychological continuum to a linear numerical index. Twenty-five

test subjects utilized such a scale and 81.8% had correlation factors in excess of 0.953 during three tests.

The use of a non-adjectival, non-ordinal scale could provide simplicity, linearity, averaging, high correlation and a high confidence for minimum testing. Such a scale, if used in contemporary testing, might greatly reduce evaluation and procurement costs.

APPENDIX A

RATER DATA

	TEST 1					
AGE	SCOPE*	SEC.	A	B	C	SCALE**
30	8	48	3	1.5	7	A
32		59	2	0.8	8	
32		57	5.25	2	5	
28		41	2	1	8	
30		56	6	2.4	4	
28		48	4	1.5	6	C
29		33	2	0.8	8	A
28		26	4	1.6	6	
27		36	1	0.5	9	
27		44	2.5	1	7.5	
29		49	7.5	3	2.5	
28	3	60	10	4	0	
28	5		7	2.75	3	

TEST 2					
SCORE*	SEC.	A	B	C	SCALE**
8	44	3	1.5	7	A
	36	1	0.4	9	
	30	5	2	5	
	24	1	0.5	9	
	42	4.5	1.75	5.5	
	52	4.5	1.75	5.5	C
	48	3	1.2	7	A
	40	5	2	5	
	34	0.5	0.25	9.5	
	36	2.5	1	7.5	
	60	9	3.5	1	
	57	10	4	0	
	42	4.5	1.5	6.5	

TEST 3					
SCORE*	SEC.	A	B	C	SCALE**
8	42	3	1.5	7	A
	39	1	0.4	9	
	33	5	2	5	
	37	1.5	0.8	8.5	
	33	3	1.2	7	
	45	4	1.5	6	C
	46	3	1.2	7	A
	39	4.5	1.7	5.5	
	33	1	0.5	9	
	60	3	1.2	7	
	39	6	2.37	4	
	55	10	4	0	
7	60	8	3	2	

*Score based on a total of eight
**Scale preference

AGE	TEST 1						TEST 2						TEST 3					
	SCORE*	SEC.	A	B	C	SCALE**	SCORE*	SEC.	A	B	C	SCALE**	SCORE*	SEC.	A	B	C	SCALE**
29	2	60	5	2	5	A	5	60	6	2.25	4	A	5	50	3.5	1.5	6.5	A
34	3		9	3.75	0.5		7		10	4	0		8	44	8	3.3	2	
29	0		9.9	3.96	0.01		3		9	3.5	1			53	7	2.75	3	
30	5		5	2	5		5		5	2	5			36	5	2	5	
29	4		5.8	2.5	4.2		6		5.2	2.3	4.8			51	4.5	1.8	5.5	
29	3		8.5	3.25	2.0		8	53	6	2.25	4		4	60	7	3	2255	
29	8	49	6	2.25	4		5		9	3.25	1		3		10	4	0	
29	0	60	10	4	0		8	59	10	4	0		5		5	2	5	
28	3		8	3	2		5	60	6	2.5	3		6		5	2	5	
29	3		4	1.75	6		4		3.5	1.25	7.5		4		3.5	1.75	7.5	
32	4		8	3	2		2		8	3	2		3		8	3	2	
32	7		5	2	5		7		5	2	5		6		5	2	5	
27	6		8	3.5	2		0		9	3.6	1.5		3		9	3.6	1.5	
30	6		7	2.75	3		2		5	2	5		3		6	2.5	4	
30	3		4	1.6	6	B	4		5	2	5	B	8		8	3.2	2	B
28	4		9	3.8	1	A	2		9.5	3.9	0.5	A	7		8	3	3	A
29	4		9	3.5	1	B	4		9	3.5	1	B	6		5	2	5	B
29	2		8	3.5	2	A	7		5	2	5	A	4		7	2.75	2.5	A

*Score based on a total of eight
**Scale preference

APPENDIX B

(Continued)

PHRASE	PSYCHOLOGICAL MEAN
Response Characteristics	
Excellent, pure (i.e., no accidental excitation) primary and secondary response characteristics	0.99
Good, relatively pure, primary and secondary response characteristics	2.47
Fair, somewhat impure primary or secondary response characteristics	4.62
Quite sensitive, sluggish or uncontrollable in primary or secondary responses	6.00
Extremely sensitive, sluggish or uncontrollable in primary or secondary responses	7.10
Effects of Deficiencies	
Effects of deficiencies on performance is easily compensated for by pilot	4.04
Some minor but annoying deficiencies	4.50
Moderately objectionable deficiencies	5.57
Major, very objectionable deficiencies	7.65
Demands on Pilot	
Completely undemanding of pilots, very relaxed and comfortable	1.65
Largely undemanding of pilots, relaxed	2.36
Mildly demanding of pilot attention, skill or effort	4.22
Demanding of pilot attention, skill or effort	5.88
Very demanding of pilot attention, skill or effort	7.50
Completely demanding of pilot attention, skill or effort	8.36

APPENDIX B

LIST OF EVALUATION PHRASES

PHRASE	PSYCHOLOGICAL MEAN
Handling Qualities	
Excellent handling qualities	1.00
Highly desirable handling qualities	1.47
Good handling qualities	2.58
Pleasant handling qualities	2.65
Fair handling qualities	4.13
Bad handling qualities	7.74
Very bad handling qualities	8.22
Control	
Extremely easy to control with excellent precision	0.97
Very easy to control with good precision	1.76
Easy to contrul with fair precision	3.21
Controllable with somewhat idadequate presision	5.43
Controllable, but only very imprecisely	6.65
Difficult to control	7.18
Very difficult to control	8.15
Nearly uncoltrollable	8.91
Precision	
Extremely easy to control with excellent precision	0.97
Very easy to control with good precision	1.76
Easy to control with fair precision	3.21
Controllable with somewhat idadequate precision	5.45
Controllable, but only very imprecisely	6.65

BIBLIOGRAPHY

1. Air Force Flight Dynamics Laboratory Technical Report 65-15, Human Pilot Dynamics in Compensatory Systems, by D. T. McRuer, July 1965.

2. National Aeronautics and Space Administration Contractor Report 1643, A Study of Relationships Between Aircraft System Performance and Pilot Ratings, by W. C. Schultz, F. D. Newell, and R. F. Whitbeck, July 1970.

3. Warner, E. P., Airplane Design, McGraw-Hill, 1936.

4. National Advisory Committee for Aeronautics, Advance Confidential Report, Requirements for Satisfactory Flying Qualities, by R. R. Gilruth, April 1941.

5. Cooper, G. E., "Understanding and Interpreting Pilot Opinion", Aeronautical Engineering Review, March 1957.

6. Air Force Aeronautical Systems Division Technical Report 61-147, In-Flight Simulation of the Lateral-Directional Handling Qualities of Entry Vehicles, by R. P. Harper, Jr., November 1961.

7. Air Force Aeronautical Systems Division Technical Report 67-19, Pilot Evaluations in a Ground Simulator of the Effects of Elevator Control System Dynamics in Fighter Aircraft, by M. L. Parrag, September 1967.

8. Advisory Group for Aerospace Research & Development Conference Proceedings No. 17, A Revised Pilot Rating Scale for the Evaluation of Handling Qualities, by R. P. Harper, and G. E. Cooper, September 1966.

9. National Aeronautics and Space Administration Technical Note D-5153, The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities, by G. E. Cooper and R. P. Harper, Jr., April 1969.

10. Air Force Flight Dynamics Laboratory Technical Report 68-76, Pilot Rating Techniques for the Estimation and Evaluation of Handling Qualities, by J. D. McDonnell, December 1968.

11. National Aeronautics and Space Administration Contract Report 1601, Motion Drive Signals for Piloted Flight Simulators, by S. F. Schmidt and Bjorn Conrad, May 1970.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
2.	Library Naval Postgraduate School Monterey, California 93940	2
3.	Chairman, Department of Aeronautics Naval Postgraduate School Monterey, California 93940	1
4.	Associate Professor Donald M. Layton Department of Aeronautics Naval Postgraduate School Monterey, California 93940	1
5.	LCDR C. Vance Schufeldt, USN Attack Squadron 212 Lemoore, California 93245	1
6.	George E. Cooper Ames Research Center Moffett Field, California 94035	1
7.	Robert P. Harper, Jr. Cornell Aeronautical Laboratory, Inc. Buffalo, New York 14221	1
8.	James D. McDonnell Systems Technology, Inc. Hawthorne, California 90250	1
9.	Bjorn Conrad Analytical Mechanics Associates Palo Alto, California 94302	1

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Naval Postgraduate School Monterey, California 93940	Unclassified
	2b. GROUP

3. REPORT TITLE

A New Approach to Pilot Rating Scales

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Master's Thesis; September 1971

5. AUTHOR(S) (First name, middle initial, last name)

Coral Vance Schufeldt

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
September 1971	46	11

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Naval Postgraduate School Monterey, California 93940

13. ABSTRACT

An examination and comparison of pilot rating scales presently in use and an investigation into the possibilities of a linear rating scale were conducted. The hypothesis was advanced that a rater may transpose his impression of performance directly to a non-adjectival, non-ordinal rating scale and thereby relate his psychological continuum to a numerical index. Experimental data, though limited, tended to support this hypothesis.

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Pilot rating scale						
Psychological continuum						
Linear numerical index						